Replies to the Critics: In Search of the Embodied, Extended, Enactive, Predictive (EEE-P) Mind.*

Andy Clark

> That the brain matches its environment is no more surprising than the matching of the two ends of a broken stick.
>
> W. Ross Ashby[1]

First of all, heartfelt thanks to Liz Irvine, Matteo Colombo, and Margarita (Mog) Stapleton: thanks for coming up with the idea of this volume, and for somehow making it actually happen. How on earth they persuaded such an amazing set of folk to contribute is quite beyond me. But I want to take this opportunity to thank them – and all my colleagues and students over an embarrassingly large number of years – for teaching me so much, and for introducing me to so many of the ideas that I have come to value and embrace.

My debts to Daniel Dennett, who so graciously contributed the Foreword, are incalculable. Dan is a lifelong friend, and the major inspiration for all my work. I am also hugely and at times painfully indebted to the many contributors to this volume.

Hugely, because these essays shed new and important light on so many topics dear to my heart – topics that may prove to be crucial anchor points for the future sciences of the mind, or just passing fancies that appear and dissolve like smoke-streams over a fire. Whatever the outcomes, I have never had so much fun, or learned quite so much, from reading a group of essays before. This is testimony, surely, to the way so many of the contributors managed to draw links between new and exciting developments in their own specialist areas and core and recurring themes in my work.

But painfully too, because along the way they reveal enough flaws, omissions, and apparent inconsistencies to keep me busy for many years to come. The pain and the gain are, perhaps inevitably, linked. The essays forced me to think harder than ever about how (if at all) the various strands in my work hang

together. Is there one picture here or many? What, if anything, holds it all together? Can I really have that many cakes and eat them all? Never one to fear bloat, I'm going to try.

I've kept the structure plain and simple. I respond briefly to each essay in turn, in a way that (or so I fondly hope) unfolds as a single, not wildly inconsistent, narrative. It's a narrative that takes us all the way from EEE (Embodied, Extended, Enactive) Cognition to EEE-P Cognition – the same core dishes, served up with a satisfying Predictive twist.

Part 1: Extensions and Alterations

The volume kicks off, appropriately enough, with an important new contribution from my partner-in-extended-crime, **David Chalmers**. Chalmers has the singular honor of being both a thoughtful defender and an incisive critic of the core arguments meant to establish that minds (human minds, as they currently exist) can sometimes extend – where that means that bio-external structures and operations become poised and woven so as to become parts of the physical machinery of thinking. In the new paper, he takes careful aim at the thesis itself, seeking a statement of the view that is neither too weak to be interesting nor too strong to be plausible. In so doing, he puts his finger on something that has bothered me for a long time.

Way back when, Daniel Dennett described a thought experiment in which a human brain was removed from the gross body, and kept in a distant location while controlling the body from afar. We might think of this, Dennett (1978, p.311) wrote "as a mere stretching of the nerves. If your brain were moved an inch over in your skull, that would not alter or impair your mind. We're simply going to make the nerves indefinitely elastic by splicing radio links into them". I have sometimes wondered whether such a scenario might be taken to establish the basic space for extended minds. The story shows that we should not be too impressed, where the machinery of minds are concerned, by the typical spatial location of brains within the bodily bounds. But at the same time (modulo very reasonable worries about time delays) there is really nothing here to challenge standard internalist initutions. Dennett's protagonist sports a mind whose core physical machinery lies clearly outside the head, but that kind of mere re-location seems deeply different to the kinds of extension defended and contested in work on the extended mind. Similarly, replacing a neuron with a radio-communicating silicon chip located outside the head seems to offer (as noted by Farkas (2012)) too weak an argument[2].

Chalmers' contribution makes clear why this should be so. The interesting thesis at issue is not that sometimes some of the machinery of mind can (in this very world) be located beyond the bounds of skin and skull. That sets the bar too low. Rather, the substantive claim is that some of the machinery of mind can safely be located beyond the intuitive bounds of perception and action. Slightly more generally, Chalmers suggests, this could be re-cast as allowing select neuro-external operations to count despite sensory and motor interactions playing a key role in enabling them to become woven into an extended circuit.

In other words, the claim is that just because some external resource (such as Otto's notebook) is subject to sensorimotor engagements, that does not erect a barrier such that all the genuinely mental activity needs to be assigned only to the 'filling' in the resulting sensorimotor sandwich. Chalmers invites me to sign up for something like this as the new official statement of the core thesis. I hereby do so, and dub it – at least for the purposes of this Reply - the 'sensorimotor liberation' story. I think it is correct, but want to float a few questions and caveats.

A potential concern is that the sensorimotor version positively invites the intervention of full-blown agentive attention within allegedly mind-extending processing loops. It does so because perception-action loops are often characterized by careful distributions of agentive attention, as when I carefully pour the rice from the packet into the narrow-necked rice jar. This can seem problematic, since there is a natural tendency to think that attention always relates a cognitive agent to some agent-external state of affairs, selecting some aspects of that state of affairs for further or more fine-grained processing. If we accept such a picture, then the intervention of full-blown agentive attention, within a perception-action cycle, seems to work (and see Clark (2015) for further discussion) against the picture of an extended cognitive process. Instead, we seem to confront a purely internal cognitive process modified by some external object, encoding, or state of affairs. It was to avoid this kind of worry that Clark and Chalmers (and then Clark (2008)) often insisted that the loopy processing be so fluent as to become almost sub-personal in nature.

The simple remedy, I think, is to notice that agentive attention can already safely intervene in purely internal processes of imagination and reasoning. Thus suppose I am wondering about the shape of someone's moustache. I call up a mental image and attend to the shape of the moustache. In such cases, we are not tempted to think of the attended information as thereby being pushed

outside the 'bounds of cognition' (to borrow a phrase from Adams and Aizawa (2001)). In the case of the moustache, I may need to attend quite carefully to my recalled image, and even then I may end up with an uncertain (hence not fully trusted) verdict. Yet this seems in no way to work against the intuitive view that that kind of fully *inner* episode is a proper part of my cognitive processing. Since we surely ought not hold extended systems to higher standards than classically inner ones, the moral is that the intervention of full-blown agentive attention is not, after all, inconsistent with the presence of extended cognitive processes partly constituted by the attended materials themselves.

Chalmers ends by leveraging the 'sensorimotor interaction' considerations so as to suggest a new reason for what many see as the uncomfortable view that the machinery of consciousness is 'in the head' even though the machinery of mind extends. The reason for the asymmetry, Chalmers speculates, may be to do with the ties between conscious experience and the 'direct availability' of information for global control. Loops through perception and action introduce way-stations such that information, while out in the loop, is not directly poised for global control. This is a fascinating idea, but one that needs a lot more development and clarification if it is to be convincing. Why, for example, couldn't currently foveated information (such as an inscription in Otto's notebook) count as directly poised? I share Dave's intuition here, but slightly longer travel routes for such signals, since they are delivered at the speed of light, do not introduce appreciable delays, so do not seem to introduce any kind of functionally relevant indirectness. Indeed, as long ago as 1972 Newell and Simon commented that "from a functional viewpoint, the STM should be defined not as an internal memory but as the combination of (1) the internal STM and (2) the part of the visual display that is in the subject's foveal view". So I think the jury remains out. It would be deeply satisfying (to me) if the 'sensorimotor liberation' rendition of the core extended mind thesis simultaneously revealed why extended consciousness is not currently actual and perhaps not humanly possible, but I am not yet convinced it does so.

Nor, with bells on, is **Fred Adams**. Adams is a long-time sceptic both about extended consciousness *and* about the arguments meant to lead us to embrace extended (non-conscious) cognition. In his characteristically friendly and forceful piece, Adams asks me to come clean on a number of issues. First and foremost he (still) wants me to offer a 'mark of the cognitive' – an account, which need not amount to a concise definition, of what makes it the case that something counts as a cognitive process at all. I have resisted this pressure on the grounds that no such account is likely to command general assent, and

because we do seem to have at least an intuitive grip on the realm of the cognitive – enough of a grip to see, for example, that digestion and photosynthesis are not cognitive processes, while in-the-head planning and episodic memory most certainly are. This is presumably because planning and episodic memory involve the repeated encoding and transformation of information into forms apt for the guidance of rational action and flexible, informed, response. This kind of rough informal grip is, I claim, is all we need to raise the question whether cognitive processes sometimes extend – a question whose answer will then help us further refine and understand the realm of the cognitive. All the old devices (most notably the parity principle that invites us to apply our informal understanding without the distractions of skin and skull) are apt for argumentative use given this rough and ready base understanding. So I continue to reject Adams' demand.

Do I therefore believe that there simply is no 'mark of the cognitive'? Adams notes that if that means there will be no principled way of ever saying of some X that X is or is not a cognitive process, that would make the very thesis of extended cognition elusive and perhaps uninteresting. But we are not working in a vacuum here – any more than moral theorists were working in a vacuum when asked to think about acceptable social orders without letting their own place in society influence their choices. Instead, the idea is that we discover how best to think about the realm of the cognitive by first liberating ourselves from a certain image of the mind as a kind of ethereal filler in a perception-action sandwich. We are then free to see it instead as a potentially looping *process* that underpins choice and action in ways that are distinctively cognitive in that they are delicately sensitive to new worldly information, and put it flexibly to use in the service of our changing goals and needs.

Adams also presses me on some passages where I suggest that cognition is to be judged more by its 'effects' than by 'causes', fearing that this hints at a kind of simple behaviorism that trivializes the suggestion that cognitive processes extend. But I didn't mean to imply that wordly behaviors might count as partly implementing a cognitive process totally irrespective of inner causes. The shape and nature of the inner parts of looping processes matter, and may be essential to the whole process counting as a cognitive process. Assuming any such constraints are met, the 'right kind of causes' – as Adams himself notes at the start of section VI - can then come to include whole perception-action loops, and need not be limited to the neuro-internal aspects of such loops. Counting on our fingers, or gesturing as part of thinking, are offered as cases like this.

Does Notto – the conjoined twin who Adams neatly imagines to play the role of Otto's notebook – constitute a reductio of the view that some of Otto's cognitive processes extend via the notebook? I don't think so. The oddity here is that Notto is a fully-fledged agent. But all that means is that two systems here know the location of MOMA – Notto (qua self-standing mind) and Otto, some of whose processing enloops Notto via sensorimotor means[3].

On a more positive note, Adams suggests that cognition "involves mental structures that rise from the level of information only to the level of meaning". I think that's right. Cognitive processes are information-transformers, that take energetic inputs and amplify, sculpt, and select them so as flexibly to serve an organism's context-dependent needs and purposes. A good way to think of the extended mind arguments is thus to see them as a way of recognizing the extraordinary extent to which those very processes of selecting, sculpting, and amplifying are realized not only by transformations carried out within the brain/CNS, but by the use of a huge range of 'epistemic actions' that likewise select, alter, and amplify energetic signals so as flexibly to serve our context-varying needs and purposes.

At this point, (certain) behaviors and cognitive processes partly coincide. But this does not mean that cognitive processes are simply *identical* with behavior. This, as **Ken Aizawa** also argues, would simply trivialize the extended mind claim itself. Instead, the message is that starting only from our intuitive grip on the realm of the cognitive, we can see (once a few skin and skull prejudices are cleared away) that many aspects of behavior, and the bio-external structures and operations that sensorimotor loops thereby poise for use, look much more aspects of the cognizing itself than like instrumental or pragmatic outflows from some purely inner cognitive machine. Where these overlaps occur, sensorimotor loops are not merely 'causal support' but help constitute the process as one that sculpts, selects, and transforms energetic information in ways that make it fit to serve our purposes.

In sum, I completely agree with both Aizawa and Adams' general insistence that cognition is not the same thing as behavior. As Aizawa points out, it is pretty darned obvious that behavior 'extends', and that cognition is meant to be something distinct – something like a certain kind of well-spring from which pragmatic behaviors might issue. But this is perfectly consistent with the idea that *sometimes*, *some aspects* of behavior are playing a recognizably cognitive role. That is the heart and soul of what I am calling Chalmer's 'sensorimotor liberation' version of the thesis of extended cognition. To insist that 'if it is

behavior it isn't (also) implementing a cognitive process' is simply to beg the question against the thesis of cognitive extension thus understood.

Aizawa also uses the cognition/behavior divide as a pivot for some rich and challenging reactions to my suggestion that public language is itself a kind of semi-external cognitive resource. The picture he offers is one in which utterances and texts help us do things we couldn't otherwise do – such as teach and learn advanced philosophy! But they do this, he argues, only by triggering thoughts 'in the head', leaving all the truly cognitive activity on the inside of the familiar inner-outer divide. There is no doubt that this is the standard picture, and that it allows the presence of public language to alter what inner cognition can achieve. Linguistic behavior is indeed behavior – so once again, the interesting claim is that some aspects of that behavior might also count as implementing genuinely cognitive processes – ones that would simply not exist were the external symbolic realm not available. Here, I direct the reader to Mike Wheeler's (this volume) discussion of how a connectionist pattern-completer, when coupled with structured external symbolic inscriptions, might implement a distinctively cognitive device – a distributed version of a 'physical symbol system' as described by Newell and Simon. This is an excellent example, since PSS's have distinctive computational properties (such as supporting symbol re-combination and 'systematicity') that seem extremely relevant to fixing the cognitive profile of an agent.

To be sure, a determined critic may still insist that only the inner elements of the processing loops here count as 'cognitive'. But this is at most a stalemate, since it is unclear what (apart from the in-the-head intuitions at issue) mandates such a view. The alternative picture is one in which some of the properties of the external medium count, despite their reliance upon sensorimotor loops, as helping to constitute the cognitive profile of the agent.

It is intriguing to note that a whole class of DeepMind systems (called 'Differentiable Neural Computers[4] ' or DNCs) fit exactly this profile, consisting of deep learning networks that have learnt to use read-write operations to couple their own internal processing capacities to stable yet modifiable external data stores so as to deliver brand new kinds of functionality. In this way, DNCs "combine the advantages of neural and computational processing by providing a neural network with read–write access to external memory…minimizing interference among memoranda and enabling long-term storage". As a result, DNCs "have the capacity to solve complex, structured tasks that are inaccessible to neural networks without external read–write memory" (both quotes from Graves et al (2016) p. 1). In short, these systems can learn to

represent and reason about complex structures – such as the London Underground system - in ways that the non-externally-augmented network cannot.

Differentiable Neural Computers are nice examples of the power of a (stripped-down) version of 'sensorimotor liberation'. They use external memory resources which they can couple with only via attentional read-write processes. These loops into external media transform the space of problems they are able to solve, delivering behavioral capacities that one would normally expect only from more classically structured problem-solving engines. I think it is at least prima facie plausible to suggest that these systems exemplify, in a minimal but revealing fashion, the way that operations made available only via sensorimotor loops might nonetheless help constitute the computational form of an embodied cognitive system. Loops like these do not merely provide triggers for inner (truly 'cognitive') operations, but look (to me) to be the material underpinnings of an extended computational process. The inner and the outer here combine in the kind of way that, in the works on language that Aizawa interrogates, I dub 'complementarity' rather than 'translation'.

Aizawa is right to note that one could, nonetheless, insist that the external stuff only does its work by having the right kinds of effect on the inner stuff. But that alone cannot make it correct to treat the inner as simply a translation of the outer. For a translation ought to have the same semantically significant properties as that which it translates. But here, structure-sensitive learning and reasoning depend crucially upon the persisting, stable, re-inspectable external store. This looks more like a distribution of cognitive labour than a simple triggering relation. Agree or disagree, I hope that the intended content of the 'complementarity' model of the role of public language is now a little easier to see. Of course, the real-world case (unlike the DeepMind system) is one in which advanced agents then come to internalize many of the initially distributed operations, making the best diagnosis less clear. For that reason, I have always thought that self-directed and self-deployed language offers an interesting kind of borderline case – a 'transition technology', perhaps, on our socio-historical path to instantiating truly extended minds.

Aizawa also questions my suggestion that public language provides a kind of 'cognitive tool' that helps us think about thinking itself. He notes that some language-less humans seem to have done quite well at thinking about their own thoughts, as shown by what they say at a later time. This is a fair challenge, and it seems very likely that simple second-order cognitive dynamics can exist without language. Language, however, may greatly enhance and extend such

capacities, enabling us to construct and comprehend chains of spiraling self-reflection that would otherwise defeat us.

Katalin Farkas illuminatingly takes up the theme of the 'extended conscious mind' (ECM) arguing that resistance to ECM is inconsistent with some of my own putative examples of extended cognition. Specifically, both the Tetris example (from the original paper) and the Ballard blocks-copying example (from Clark (2008a)) seem to involve events that are 'part of the stream of consciousness' – unlike the merely standing beliefs of Otto. However, a conscious moment can be part of an extended mental process even if the conscious moment itself is wholly internally supported – such, roughly speaking, was the intended upshot of the Tetris and Ballard cases.

Still, it seems fair to ask why the cake be thus carved. It is not (for me or, I think, for Chalmers) because conscious processing resists functional specification. Rather, it is because – for whatever reason - inner processing seems sufficient to deliver the conscious state at every moment. Conscious experiences of the Tetris screen might thus all be constructed 'in the head' (just as Farkas, in section 5, insists). So what remains to count as an extended cognitive process? The idea was that those conscious moments, internally constructed, are part of larger, extended, genuinely cognitive processes. For example, to get the *right* conscious state at the *right moment*, certain loops into the world (for example, manipulating the zoids so as to aid identification) may play a crucial role. Indeed, I would expect that a lot of the relevant Tetris-playing actions are launched and initiated without conscious involvement at all! In any case (to take up the kind of story about conscious experience suggested by Chalmers in his contribution) we must surely allow that non-conscious *inner* processing can play a crucial role in the run-up that poises information directly for the global control of action and reason. But then by the same token, there is no reason why (if the extended mind arguments are on track) non-conscious elements of the bio-external flow cannot play the same kind of role. In each case a behavior (zoid placement) may be proximally caused by the conscious state, even though that state arises as part of a process that may potentially extend, and that is not to be identified with its conscious moments alone. I recognize, however, that treating conscious experience as a succession of 'moments', rather than a rolling process, seems deeply wrong (and see Clark (2009b)). As a result I am increasingly open to the idea of cognitive extension for at least some conscious states.

Farkas is exactly right to insist that cases of simple 'external re-routing' (or 'cognitive prosthesis') are not properly speaking cases of extended cognitive

processes. Such cases (which would include the 'mere stretching of the nerves' imagined by Dennett) are at best useful softeners, reminding us that location is not functionally essential to any computational process. But they are not exemplars of the kind of deep conceptual challenge attempted by core arguments for the extended mind. The difference, as Chalmers (this volume) rightly suggests, is that the core arguments all seek to establish the real-world possibility of cognitive extensions that involve bio-external operations made available via sensorimotor loops.

Michelle Maiese usefully canvasses a number of reasons to doubt that *affective states* extend. Her principal target is work by Colombetti and Roberts (2015). Here, I tend to agree (though with increasing uncertainty – see above) with the general conclusion, but was unconvinced by the argumentative route on offer. Maiese (this volume, ms page 5) suggests that "crucial structural aspects of emotion –such as its egocentric, spatial, and temporal dimensions—are physically grounded in the neurobiological dynamics of living organisms… and [that] this lends support to the thesis that emotional consciousness is constitutively dependent on our living bodies". But this is in tension, she argues, with appeals to arguments for the extended mind insofar as they imply belief in the thesis of multiple realizability, and thus allow that an a bodiless brain-in-a-vat could enjoy all the same experiences as we do. The idea is thus that Colombetti and Roberts ought not to make their case for extended affect by appeal to (versions of) arguments for the extended mind.

I want to resist the suggestion that accepting a brain-in-a-vat (BIV) scenario implies that "the specific details of human embodiment would play no essential role in cognition" and that a bodiless being might enjoy all the same mental states as we do. I resist this because I do not believe the BIV agents to be bodiless at all! Following Chalmers (2005) it seems to me that the BIV picture is best understood as one in which the body itself, with all its cognitively relevant physical idiosyncrasies, is alternatively 'deep physically realized'! BIV agents, on this view, are as genuinely embodied as you and I, and their bodies have specific, potentially cognitively-relevant, forms (see Clark (2005), (2008a)). It is just that the deep physical bedrock of those bodily forms is not quite as we normally imagine it to be.

Let's call this kind of alternative deep physics 'alt-physics'. Next, consider that the BIV agent, despite her realization in alt-physics, may still count on her fingers, and will be as impacted as we are by the number of fingers readily available. Like us, she will exhibit (Maiese ms p.5) "racing hearts, quickened breathing, grimacing faces, tensing muscles, tingling skin, and sweating palms"

as well as "alterations in skin conductance, cardio-pulmonary changes, and musculoskeletal changes". She may also leave sticky notes on her desk, and these will be subject to the same disruptive physical forces (wind, rain, other agents) as when realized by more 'standard' means.

All this is consistent with my view (more on which below) that sometimes, the very same mental state might exist in differently embodied or embedded agents thanks to some compensatory adjustments in the distribution of labor between body and world. To see this, notice that such cases would *themselves* be apt for full re-creation using 'alt-physics' as the bedrock. In such a case there would be two differently alt-embodied but mentally identical agents realized using alt-physics. Bodily differences, though often cognitively relevant, need not always be. Bodily difference often matters, and it often makes a cognitive difference. But bodily difference is not sufficient (see Clark (2008b)) for mental or cognitive difference. Alt-physics, on the other hand, is not cognitively relevant at all.

These themes also animate **Larry Shapiro's** lively and compelling exploration of the various ways that embodiment might matter for mind. Shapiro resists my (2008b) arguments supposed to favour a 'larger mechanism' account of extended cognition over a 'special contribution' account. 'Larger mechanism' (LM) stories would include the kind of case mentioned above, where differing bodily forms share mental states in virtue of compensatory adjustments in the overall balance between bodily, neural, and worldly contributions. 'Special mechanism' (SM) stories stress the many unexpected ways in which specific details of embodiment impact mental states.

Shapiro agrees that sometimes, compensatory adjustments might occur, and that the LM (larger mechanism) story is viable. Differently embodied beings, when viewed through 'psych-goggles' – imaginary spectacles that "filter from view everything but computational structure" – might thus look exactly the same. But he thinks that SM is more interesting for psychology and cognitive science, and that LM (as neatly dramatized by the psych-goggles) even threatens to blind us to the importance of real human bodies for human mental states.

Shapiro's example concerns the way right- and left- handed people differ in some judgments, tending (after controlling for content) to positively favour options and choices whose descriptions are presented on the dominant side – on the right, for right handers, the left for left handers. Commenting on this result, Shapiro suggests that whereas LM might here reveal a computational

commonality, SC would reveal "why, because of their differences, right- and left-handers will display differences in their cognitive propensities." Shapiro glosses this by saying that "the focus of SC is squarely on cognition, not implementation, and the role of the body in shaping or informing cognition".

LM thus threatens to focus attention away from the very stuff (the body and world) that it seeks to celebrate! SM, by contrast, is free to revel in the many ways embodiment 'permeates psychology'. However, I think this is a false choice. Rather than thinking simply in terms of a 'level of algorithm' and a 'level of implementation', we do better to think of many levels such that what is psychologically interesting varies with explanatory project. For example, there are various algorithms that sort numbers into order. For some purposes, the differences don't matter. For others (e.g. understanding the relative time it takes to carry out different sized sorts) they matter a lot. Similarly, I do not think there is a single 'level' of description that best serves psychology. In the end, my arguments for LM are perhaps best seen in that light. In the case of the right and left handers, there is a psychological profile that differs and one that does not. Both – just as Shapiro himself ends by suggesting – are important if we are to unravel the importance of embodiment to the human mind.

Mike Wheeler addresses core issues concerning how best to argue for the extended mind. Like Fred Adams, Wheeler believes that extended mind debates should revolve around some agreed 'mark of the mental'. His arguments fit neatly with our earlier discussion of Differentiable Neural Computers. Briefly put, the suggestion is that appeals to hybrid connectionist systems that also manipulate external symbols (as discussed by Bechtel (1994, 1996)) or to Differentiable Neural Computers (in contemporary work on Deep Learning) will tend to favor a 'merely embedded' rather than a truly extended picture. This is because the 'rough folk intuitions' that I myself appeal include (Wheeler argues) the intuition that the real machinery of mind is in the head. The way out, he suggests, is to start instead with the principled appeal to an independently motivated mark of the cognitive, such as might be provided by Newell and Simon's Physical Symbol System Hypothesis (PSS). For it is visibly the case that it is only the combined inner/outer machinery that (in the cases just mentioned) provides for the kinds of operations upon stable chunky symbols that PSS itself requires. Complementarity of the inner and outer contributions here ensures that the hybrid inner-outer system itself is the organization that displays the mark of the cognitive.

Wheeler is right that IF processing in accordance with the PSS were accepted, in anything like its original form, as a mark of the cognitive, then the extended

(but not the non-extended) engines would here qualify. But what this is bound to suggest, to e.g. connectionists with more internalist leanings, is simply that the PSS is not the mark of the cognitive after all. Most likely, a theorist such as Bechtel will simply conclude that something else (powerful pattern-completing abilities perhaps) picks out the truly cognitive, with greater behavioral success being pressed from those resources by their canny couplings with external props and aids. It is for this reason that I still believe that the parity arguments, combined with rough folk intuition, are the best way forward. For even if the folk tend to think that the cognitive is all in the head, the parity arguments invite them to put that element of the folk picture aside and to reconsider their own intuitions without that singular distraction.

To be honest, however, I no longer believe that any side can 'win' these debates concerning productive embedding versus true cognitive extension. Instead, what those debates seem to me to have revealed is deep uncertainty or conflict within both the folk and cognitive scientific understandings of mind and cognition themselves[5]. It is depressingly clear that a significant part of the recent history of Philosophy of Mind and. Cognitive Science has been devoted to as-yet-unsolved debates concerning the applicability, or otherwise, of standard mental predicates to a variety of systems, organisms and processes. Disputed territory includes thermostats (Dennett (1987, 1998)), paramycia (Fodor (1986)), language-less animals (McDowell (1994)), 'swampmen' (Davidson (1987), programs (Searle (1980)), plants (Trewawas (2003), Calvo and Friston (2017) and certain forms of putative sub-personal cognitive activity (Searle 1992)). What this shows is that there is simply no easy consensus among 'suitably trained observers' (folk or otherwise) concerning the distribution of minds and mentality in either our natural or technological worlds. This lack of agreement is equally evident in daily life, when one considers debates over other animals, fetuses, pre-linguistic infants, some coma patients, and so on. Our unresolved collective uncertainties concerning putative cases of cognitive extension are perhaps less surprising when seen as part of this larger pattern.

I continue to believe that radical internalism is scientifically unjustified and unjustifiable. Yet the radical externalist option (extending the mind by sensorimotor means) is inevitably revisionary, and apparently sits ill – just as Wheeler argues - with deep folk intuitions concerning the biological 'innerness' of mind, perhaps due to the proximity of those folk notions of mentality to notions of (the machinery of) conscious experience. Despite this standoff, I'm optimistic about the overall impact of these debates upon science, philosophy, and even, as time goes by, upon common-sense. For what is now very widely agreed is that it is the complex weave of inner and outer capacities, tricks, and

ploys that together make possible the varieties of human and (other) animal thought and adaptive success. Within those weaves, there are myriad previously unexpected contributions made, as Shapiro and others suggest, by gross physical features and by sensorimotor loops that create and exploit bio-external order. Understanding this rich and surprising mosaic, however described, is the real task of the sciences of mind and behavior.

Part 2: On Being a Cyborg

Part 2 opens with a characteristically probing contribution from **Louise Barrett**, pitched from the productive crossroads between anthropology, psychology, and (the rest of) cognitive science. Barrett's use of work on embodied and extended cognition as a route towards a better understanding of non-human animals is striking proof of the practical value of these new perspectives. Seeing beyond the cognitivist brain-bound paradigm enables us, Barrett has argued, to avoid anthropocentrism as we consider other species – it invites us to appreciate their kind of cognition, bound up with their kinds of bodies and their kinds of environmental niche.

Barrett is unconvinced, however, by my ongoing use of the 'cognitivist' vocabulary of internal models and representations, action-oriented or otherwise. That worry is potentially exacerbated, she notes, by my more recent appeals to generative-model based prediction machinery (in the context of work on 'Predictive Processing' or PP for short) as lying at the very core of flexible adaptive response in both human and non-human animals. In that work, my long-standing use of terms like 'action-oriented representation' and 'internal model' is joined by talk of multi-level probabilistic inference. But is all this cognitivist-sounding talk useful or justifiable? The issue is personally pressing, since I simultaneously argue (Clark (2016a)) for a somewhat revised understanding of all these key terms – one that reveals the prediction engine as fundamentally tuned to affordances and geared to the use of body, world, and action as means of simplifying inner processing. So why not just swallow the ecological medicine and ditch those cognitivist descriptors once and for all, replacing them with talk of affordances, tunings, resonances and the like?

There are several reasons (see Clark (2016a)) why I am not (yet) persuaded to do so. But most fundamentally it seems to me that doing so obscures useful information about what happens during both learning and ongoing response. If these stories are correct, learning involves the gradual installation, in the human

brain, of a multi-level generative model defined over sensory outcomes. To be generative, in this sense, is to be capable of constructing plausible versions of those sensory outcomes using prior knowledge ('from the top-down'). That knowledge ends up being a highly structured resource, that locks on to patterns at many scales of space and time, in ways informed by the nesting of actual causes in the world.

For example, an action may be predicted both at a gross level ('get the roasted red pepper and hummus sandwich') and at multiple related lower levels, slowly unpacking the goal into a set of smaller action-sequences, that are in turn cashed out via predictions about the required states of muscles and tendons. Ditto for simply seeing that scene, where predictions about the general nature of the setting ('in a vegetarian-friendly sandwich-bar') inform predictions about the items on offer, that inform predictions about their look, feel, and taste. Not only that – the system becomes able to generate all these outcomes in ways that must take account of a wide variety of contextual cues, including interoceptive cues involving our own bodily states. The generative process is thus delicately responsive to inner and outer context, spans perception and action, and is defined over a rich space informed by what might most naturally be cast as 'representations' of nesting, interacting worldly and bodily causes operating at multiple interlocking scales of space and time.

How should the representational skeptic capture this picture? Bruineberg and Rietveld (2014) suggest we speak instead of an system apt to support a good, affordance-based 'grip' on the world. But I suggest that the most informative word remains simply 'representation', but now stripped of all connotations involving chunky symbolic items of the 'Language of Thought' variety. The picture of ongoing attempts at prediction as installing a probabilistic generative model that *represents* both actions and states of affairs is simply a way of drawing attention to the fact that (1) the generative model is here meant to be a real aspect of our cognitive organization (not just a theorist's fiction), and (2) that it is takes the form of a structured and flexible knowledge base, apt for repeated redeployment in ways that are appropriately sensitive to new information.

Might we do sufficient justice to the highly structured nature of that knowledge-base by describing it simply as being *tuned to* or *resonating to* the right stuff at the right time? Maybe – if we now understand the notions of resonance and tuning as involving the separation of causes into nesting factors operating at different but interlocking time-scales, and stress that the resonating reflects inner and outer context and is appropriately sensitive, again at multiple time-scales, to new information. But thus refined these notions of resonance, tuning,

and grip look (to me at least) to be just action-oriented representations called by another name. Perhaps that means we all agree on the substantial claims hereabouts, and can finally call an end (Clark (2015)) to the 'representation wars'?

Barrett also worries that appeals to internal representation make a kind of tacit but illegitimate reference to some kind of inner user or interpreter – the rightly-feared inner homunculus. However, the cognitive scientific use of 'internal representation' was from the outset meant to be thoroughly purged of that connotation. Barrett is right, however, to note that this means that the generative model is not *representing the world to the agent*. Rather, it functions *within* the agent to enable apt response to the world. But better yet, the prediction error signal here provides the perfect means of driving multi-level multi-scale processes of self-organization. For prediction error is a self-computable signal that enables both learning and ongoing online response. By stressing self-organizing around prediction error, every hint of the controlling homunculus is purged from the PP paradigm.

Rob Goldstone invites us to consider something that he calls the 'Reverse Parity Principle' (RPP). RPP states that " a brain component should be considered to be part of a distributed cognitive system if we would accept it as being part of a distributed system if it were non-biological." This is consistent with the idea that sometimes, individual cognitive systems extend and is another manifestation (as Goldstone notes) of the idea that we can profit from thinking about complex functional organizations in ways that are not biased by metabolic boundaries. Applying this in the opposite direction to the extended mind corpus, RPP allows us to take intuitions developed by looking at distributed organizations and apply them to the organization of a single brain.

This is a neat flip, and one that strikes me as both legitimate and illuminating. In particular, Goldstone stresses three features that have been explored in distributed settings and that might be re-applied to the organization of a single brain. The features are specialization, tool creation, and 'indirect levers'. Specialization within distributed organizational forms is manifest in the division of labour between elements and groups. But within the brain, it speaks to the increase of internal structure over time, and the developmental emergence of functional specializations in some brain areas even for evolutionarily recent developments such as reading and exact mathematics. Tool creation means what it says, the creation of new tools for tasks, including tools for toolbuilding. But applied in the single brain setting, it suggests the creation of new mini-systems that help us do specific kinds of things better. This potentially overlaps

with 'specialization' but Goldstone especially stresses the role of 'perception and action modules as tools shaped – by automatic unconscious processes - to our needs'. Finally, 'indirect levers' are ways of shaping our own responses that can work even though we are unable to reach directly into the brain to do so. Instead, we might 'hack our own minds' by means of novel targeted training methods, some impressive examples of which are given in the text.

Goldstone's larger vision here looks to be one in which the brain is seen as a bag of tricks, but a bag whose nature may be best understood by looking at some general principles and tactics whose operation is visible in many kinds of organization, both inner and outer/distributed. I am sympathetic to this broad vision, but suggest that another (not incompatible) key organizing principle involves the organismically self-computed 'prediction error signal'. This quantity, as argued above, arguably plays a key role in both learning and ongoing response. Most importantly for present purposes, it enables self-estimated uncertainty to *orchestrate* – without the need for an overseeing homunculus – the activation and use of the right inner resources at the right time. In so doing, it may also recruits motor routines that can fold in the use of gross bodily or bio-external resources as required, deferring to those external resources whenever that reflects best overall policy in problem-solving context. In this way, self-organization around prediction error may be the glue that binds inner and outer into coherent but transient ensembles.

David Kirsh asks – very reasonably in my view – what is really at stake in the debate between those who embrace extended cognition and those who see the mind as 'merely' productively embedded in ways that make good use of the body and world? Kirsh anchors his discussion in (pretty robust) intuitions concerning the status of physical prostheses as candidate body-parts, and offers four criteria to distinguish the extended and the merely embedded. The first two are unproblematic. True extensions, Kirsh suggests, should be temporally tightly coupled to, and 'in sync with' with the rest of the system – just like a good prosthetic limb. And the coupling needs to be phenomenologically transparent at the time of use, so it serves our wider purposes rather than itself being an object of thought.

A third requirement is that the bio-external stuff be orchestrated and controlled by the person's own inner processes. Thus, I should control my prosthetic limb, rather than it act alone or so as to control me. Taken as a general rule this one strikes me as potentially more problematic, since it suggests the kind of asymmetry often used to motivate quite strongly internalist viewpoints. What seems right is the idea that whichever bit or bits of an extended bio-

technological hybrid system happens to be currently 'in control', the overall phenomenology is one of fluid agentive doing. Just as in skilled driving an automatic braking system (ABS) can autonomously act so as to balance the braking action, so control might sometimes be located in the bio-external parts of an extended system. In a futuristic prosthetic limb, perhaps the fingers, or other brand-new effector forms, might make some local decisions in a similar kind of way to the ABS?

Finally, Kirsh suggests that we should require 'bidirectional coupling', so that each element can materially affect the other. This helps avoid trivial-seeming extensions involving 'just looking' at a target say. I agree that these cases are not thereby rendered ones of cognitive extension. But I am not convinced of the requirement. A brain area that merely stores information that other areas can later access, without those acts of access altering the stored information, would surely have counted as part of the cognitive system. Even if no such areas exist, it seems wrong to rule them out merely on grounds of failing this form of bidirectionality. What really matters, I suggest, is just that the area or resource is one that is reliably invoked as part of a process which, were it all to occur inside the head, would uncritically be accepted as a cognitive process of that agent. There is no obvious way to then render 'just-looking' cases acceptable, since that would mean paradoxically nesting the actual environment (rather than some kind of trace or representation of it) inside the putative agent!

Kirsh then looks back at the original Clark and Chalmers case studies, arguing that the Tetris case (which as based in part on work by Kirsh and Maglio (1994)) meets the revised, prosthetics-inspired set of requirements but that the infamous Otto case does not. This is because external rotation is tightly coupled in all the right ways, whereas consulting the notebook fails to offer the kind of tight, fluid, integration seen in Tetris or prosthetic limbs. The picture here is one in which consulting the notebook is a rather hit and miss process, temporally patchy, subject to multiple kinds of breakdown. The combined upshot is that it fails to become transparent equipment. In response, I'd note that Clark and Chalmers do insist that the notebook share key aspects of the functional poise of bio-memory – that it be fluidly and automatically deployed as and when required. So anything that gets fundamentally in the way of that simply violates the conditions of the original thought-experiment. The hit and miss scenarios are thus ruled out without needing to endorse any (to my mind somewhat problematic) further constraints on cognitive extensions as such.

What about bio-external resources when they are not in use? The picture that Kirsh offers – one whose phenomenal requirements I fully endorse – applies only at the moment of tightly-coupled use (the moment of automatic systemic invocation). Concerning such resources at other times, Kirsh' suggestion is that this becomes a matter for law and precedent rather than philosophy and science. In one way this seems right. But there may be fundamental principles we can appeal to as well. For example, I suspect that a lot will turn on the ease of replacement of the contested resource. If I can very easily get a new prosthetic that fits just as well, it may not count strongly as part of me when not in use. But in fact, prosthetic limbs take time to become well-fitted and it can be quite traumatic when new ones, even of the very same model, are required. It may be this personalized dovetailing (rather than ongoing spatial location or ongoing active coupling) that is the key feature. If so, then delicately dovetailed, hard-to-replace stuff that meets conditions for cognitive incorporation should count as part of 'my cognitive apparatus' even when not actively in use. After all, it may be that aspects of my internal cognitive system are sometimes not 'in use' (perhaps during deep sleep).

Returning to that compelling prosthetic metaphor, the 10th century Persian philosopher-scientist Avicenna (Ibn Sina) observed[6] of our own biological body-parts that:

> "These bodily members are, as it were, no more than garments; which, because they have been attached to us for a long time, we think are us, or parts of us [and] the cause of this is the long period of adherence: we are accustomed to remove clothes and to throw them down, which we are entirely unaccustomed to do with our bodily members".

I leave the reader to conduct the interesting exercise of repeating this line of reasoning for aspects of our own internal cognitive apparatus not actively 'in use'. (Hint: let's not risk shrinking the mind and self to the size of a boot program – a few lines of code that get the operating system up and running).

Kim Sterelny sets out in lively pursuit of the extended mind 'over deep time'. Sterelny's main focus is on the emergence and explosion of material and cultural scaffoldings – what made that possible, and when that vastly empowering and transformative incremental snowball really got rolling. At the heart of his story lies a fascinating puzzle. There is a large temporal gap, it seems, separating the first appearance of 'anatomically modern' humans and the explosion of material culture. Sterelny, after discussing the possible role of demography and social organization, turns to a very interesting further

(consistent and complementary) possibility. This is the possibility (neatly defended by Heyes (2018)) that cultural evolution itself delivered innovations that streamlined and improved cultural learning, building mind-tools upon mind-tools (cognitive gadgets upon cognitive gadgets) so as to enable the explosion of material culture itself. According to Heyes, key capacities such as 'mind-reading', language, and imitation learning may themselves be cultural products – software updates installed not by evolution but by general-purpose learning mechanisms. But once those updates are available, the snowballing process of incremental cultural learning and transmission really gets going. For some of the updates are updates that permit the sharing, improvement, and transmission of updates. If I read him right, Sterelyn thinks that further key aspects of this socio-cultural 'ignition' may be documented by Kelly (2015) in important work on formal performances, mnemonic devices, ritual, and oral traditions. Unfortunately, 'soft' technologies leave few traces so precise time-lining remains, Sterelny argues, problematic.

That said, the general picture is compelling. Our extended and (more generally) highly socio-culturally and materially scaffolded minds are not the direct products of key genetic variations in our species. This is quite unlike, say, the remarkable mole cricket, some species of which build and exploit physically perfect 'Klipsch horns' as a means of efficient sound production. The crickets – beautifully described in Turner (2000) - benefit from genetic programs that lead to feedback-tuned excavation and tunneling. With the ground itself then acting as a very large (near-infinite) baffle, and the boost of a (sometimes double) exponential horn, these crickets obtain an almost unimaginable increase in acoustic efficiency! The cricket thus benefits from bio-external scaffolding. But that scaffolding is just another expression of their bedrock genetic organization. Our human tools and scaffoldings, by contrast, seem open-ended and self-multiplying: in our worlds, even the *strategies needed* to build specific external structures (such as paper-mills, universities, and factories) can be preserved, improved, and transmitted by non-biological means. Minds like ours are thus open-endedly extensible, with extensions breeding new extensions, and no limit currently in sight.

Section 3: Embodied, Extended, but Predictive Too?

Mike Anderson and Tony Chemero speak up strongly for the role of *ecological information* (which I'll expand upon shortly) in the construction of adaptive response. Getting straight about ecological information, they argue, shows how we can be truly *in touch* with the world, even if we are also (as work

on 'predictive processing' (PP) claims) in the business of predicting the shape and evolution of the impinging sensory flux. Doing so helps reconcile PP with the full gamut of insights from embodied cognition, as argued in Clark (2016a)). But it also forces us (Anderson and Chemero suggest) to adopt a more deflationary reading of key aspects of the PP apparatus itself – specifically, a deflationary reading of PP's appeal to inner models, information, and representation.

The paper start with a great example – the way we humans sway when standing. That swaying, they report, is not some kind of motor defect so much as a functional strategy that helps us harvest good information for the control of adaptive response. It turns out that the form and directionality of the swaying alters in task-sensitive ways, and is best understood as a means of generating 'optic flow' that reveals task-relevant features of the world.

The brain's role in such a process is well captured by PP. I'd gloss that 'capturing' by saying that PP co-constructs perception and action, and that those bodily movements (the functional, context-sensitive swayings) reflect neural best-guesses about where, when, and how task-relevant information is to be found. Anderson and Chemero construct the case in a subtly different way. They go on to suggest that the ecological perspective offers a direct route into 'semantics' here. They say:

> "Ecological information—the information available to a moving animal in the environment—is *inherently semantic* because it specifies the affordances of that environment, what the animal can do in that environment, and generates and supports expectations for what that moving animal will experience as it moves. Ecological information reveals the world as significant for a given creature." (Ms page 4, my emphasis)

I quote the passage in full because I pondered it, in its full textual context, for a very long time. It seems to suggest (in context) that the question of how minds get to 'make worlds available' is resolved by noting that some systems get to pick up on the affordances that matter for the kind of being they are. A lot here depends on what it means to make a world available. A simple robot could do this kind of thing. It would properly be assigned an Umwelt. But the simple affordance-sensitive robot need not thereby experience any world at all.

A few paragraphs later, Anderson and Chemero press their concerns further, suggesting that common constructions of PP make the mistake of down-

playing the richness and potency of all that ecological information, yielding a spurious puzzle (how to deal with an impoverished signal) whose solution then appeals to inner models as a way of adding back the missing stuff.

I think their idea is this. If you start (Shannon-style) from mere non-semantic 'information' you will never bridge the gap between mere information and semantics. But if you start from the enriched land of 'ecological information', there is no gap to cross in the first place. That's because ecological information directly structures apt action, and apt action is in some sense all there ever is to meaning. This secures a strong form of epistemic directness - we are in touch with a world as it matters for our needs and purposes. This has a superficially very different flavor to saying that our access goes via some inner predictive model of the world. Surely a view that depicts perception as a highly constructive, model-driven process threatens to cut us off from the world that ecological information makes so readily available? So we seem to face a stark choice: perception as ecological pick-up, putting us in touch with the world, versus perception as model-based 'controlled hallucination', keeping the world at arm's (or maybe mind's) length. The lurking threat is that PP, constructed in the latter way, delivers us straight back into the arms of Descartes – ethereal minds cut off[7] from the world, encountering only our own 'controlled hallucinations' or 'virtual realities'. The appeal to ecological information, by contrast, works against the idea that the worldly information is impoverished and in need of sifting and enriching by inner models.

But perhaps the debate should not be structured around these old constructs of directness and indirectness? An alternative view (and the one I tried to suggest in Clark (2016a) and elsewhere) is that PP provides promising apparatus for dissolving the tension, by allowing a system to rely to a greater or lesser extent on ecological information according to changing contexts and task-demands. Prediction error here appears as the prime principle of self-organization. If task-salient prediction error is easily resolved at lower levels (as with simple reflexes and over-learnt behaviors), a system will look to simply be tuned to its environment, in ways that seem essentially model-free. But if prediction error is pushed further and further through the system, recruiting complex world-knowledge and requiring effortful processing, it will look to be functioning within a richly model-based economy. The truth of the matter, though, may be that the core goal of maintaining behavioral success is always and everywhere achieved in the same general fashion, by using estimated precision to recruit and emphasize just those aspects of the available information, and to recruit and temporarily privilege just those inner and outer resources, that working together offer the most frugal solution to the creature's ongoing needs and

long-term goals. Huge swathes of work in contemporary PP are devoted to showing, in quite fine detail, just how these checks and balances may operate, allowing creatures to balance exploration against exploitation, epistemic foraging against pragmatic action, and short-term interests against long-term plans.

Does that picture leave us cut off from the world? I don't think so. The PP strategy of using self-estimated uncertainty to repeatedly reconfigure complex organizations so as to reduce task-salient prediction error allows these systems to make maximal use of ecological (often self-generated) information, within a cognitive regime that can sometimes look to be operating almost model-free, at other times model-rich, and (crucially) *all points in between*. There is a sense in which perception, especially when it involves complex predictions from the higher levels, could indeed be thought of as 'controlled hallucination'. But PP also shows just how perception strips away unhelpful noise to reveal the core shape of the task-salient signal – a signal frequently apt to guide behavior in just the way the ecological frameworks suggest.

Anderson and Chemero close with the pregnant suggestion that standard philosophical conceptions of belief, knowledge, and justification maytransformed beyond recognition once we take a thoroughly 'affordance and action-oriented' picture on board. They are right. Re-modeling the old epistemology in ways that respect the fundamentally action-centric picture emerging from contemporary cognitive science is, I think, one of the most pressing tasks for 21st century Philosophy of Mind. But my guess is that such re-modeling will reveal these questions of 'inferential seclusion' versus 'open access to the world' as pseudo-questions that depend for their very intelligibility on mistaken pictures of what it means to perceive and to know a world.

Karl Friston – whose patience with a non-mathematically-sophisticated philosopher never ceases to delight and amaze me - urges me to 'take the high road'. As the son of a Scotsman, the 'low road' to me usually means death (considered, in the famous refrain[8], to provide the fastest way back to Scotland) while the high road consists in the normal terrestrial route! But Karl's high road is something rather more exalted – the derivation of PP itself from first principles involving existence, persistence, ergodicity and Markov blankets.

In brief, the high road starts by noting that creature's that exist do so by preferentially inhabiting the states (inner and outer) that are necessary for them to do so. This attracting set of states makes them mathematically 'ergodic', visiting and re-visiting the very states that define them as the creature they

happen to be. This, in turn, makes sense only if there is a demarcation of some kind between the creature (or system of interest) and the rest of the universe. Such demarcation is captured by the technical notion of a Markov blanket, here, the set of states forming the boundary at issue. Now we have a blanketed system that seems (from a certain perspective) temporarily to resist entropy, using up energy to stay within the very states that define it. Any such system, Friston observes, is a model of its world in – but perhaps only in - the sense that it will seem, in its behavior, to be in the business of locating the good (viability- maintaining) states/places and avoiding the others. This, Friston notes, is as true for for a bacterium as for a bishop. Indeed, we might even see the humble oil-drop (op cit p.4) in the same broad way. All these systems persist because (tautologically) they act and react in ways that preferentially harvest the kinds of state that define them, thus maximizing their own 'self-evidence'. In this way "the very fact you exist means that you will behave like you have a model of your world that predicts sensory samples with a high accuracy and minimal complexity" (Friston, this volume, ms page 5). The free energy principle itself, Friston then notes, amounts to a re-statement of these tautologies and implicatures. For it is is "just a way of defining systems (Markov blankets) that exist (are ergodic).". This is what Friston aptly describes as "the tautological denouement of the high road."

I draw the reader's attention to the use of language hereabouts. Friston says that such systems will behave 'like' they have a model of the world. Elsewhere in the same text, there is talk of them as behaving " they have beliefs about the world. We also read that a model, in this usage, is "just an ergodic system or phenotype", reminding me of his response to my (2013) BBS paper, in which he argues that agents don't *have* predictive world-models, but simply *are* such models This is all quite revealing. Talk of models, in this sense, looks to be computationally and ontologically undemanding (just as Anderson and Chemero argued). It is perhaps best understood[9] simply as a handy use of language by the scientist (the modeler) rather than a substantial claim about the cognitive organization of the modeled system. I'll return to such matters shortly.

The low road, meanwhile, is not – I'm glad to report - death but the 'effective information processing' route to PP. That route is pursued at length in Clark (2013) (2016a) and elsewhere, and invokes primarily consideration of bandwidth, speed, and flexibility. By using cascades of top-down generative-model based prediction, and feeding only residual errors forward level by level, creatures like us are able to devote expensive processing resources only to what is newsworthy in the sensory stream. And by repeatedly reconfiguring our own processing in ways guided by self-estimated uncertainty ('precision'), we are

able to respond flexibly to changing tasks and needs, sorting signal from noise in ways that are deeply (inner and outer) context-respecting. This is meant as a substantial and falsifiable claim about the cognitive organization of the target system.

Friston argues that this low-road landscape can, and should, be derived using only the high road (sparse, tautological) ingredients. I'm not really sure about that, one way or the other. But even supposing that it is true, it does not follow that all the low road talk of PP systems as having or deploying (and not simply being) multi-level probabilistic models of their worlds, or as representing those worlds in appropriately action-oriented ways, or as approximating Bayesian inference, is thereby revealed as demanding no more than the same talk does when applied – via the high road - to the bacterium or the oil drop. The high road delivers notions of prediction, model, and Bayesian inference only in highly attenuated ('as if') form, whereas the low road – or so I claim - amounts to a very substantial empirical bet concerning the organization of information flow in the brain/CNS, and the way that flow contributes to embodied agency. So much as I admire the inviolable beauty of high road, I prefer to keep it at a safe distance. It may well offer a set of fundamentally tautologous principles that frame the entirety of life and mind (even if the oil drop remains something of a puzzle). But that strength is also a kind of liability, for it means that evidence that every actively self-maintaining system conforms to the free energy principle and displays the high road profile is hardly news. They have to, on pain of failing to persist as objects of study at all. What is left wide open is the question of how different systems do so, which is the home of various possible process stories of which PP is just one among many.

Finally, I note with interest that the Epilogue to Friston's piece iterates the high road considerations, turning them back on themselves so that some creatures will look as if they have beliefs about their own generative models, in ways that might lead them to display behaviors such as novelty-seeking and the pursuit of science and philosophy. Karl then asks me if I believe – given such tempting high road meta-constructions – that there are any 'imperatives that live beyond the free energy principle'. I'm not sure, perhaps because I'm not sure what it takes to be an imperative in this sense. But I do think that we need to give processes of incremental cultural change some real credit hereabouts. A leading lesson of the work on embodied and EEE cognition is that we repeatedly build new social and material worlds around ourselves, including new worlds to train ourselves to think about our worlds. It may be only courtesy of these slowly culturally-installed lenses that there ever emerge complex systems that not only predict worlds, each other, and sensory outcomes but that *also* model

themselves as model-using systems that predict worlds, each other, and sensory outcomes.

Next up, **Jakob Hohwy**[10] continues his engaging, friendly, yet philosophically incisive campaign to push me out of my comfort zone, using PP to argue for a picture that involves "rich and reconstructive, detached, truth-seeking inner representations, characterized by fragile inferential processes, and harboured within the nervous system's unitary, fixed, non-extended Markov blanket" (ms p. 1). This – as should be clear from the previous responses - is not the mind-world relation that I think PP delivers.

One key place where we part company, it now seems, is in our conception of the role and functioning of the core 'precision-weighting' tool itself. Precision-weighting, recall, reflects self-estimated uncertainty and alters the influence of specific predictions or prediction errors on ongoing processing. High-weighted prediction errors, to take the basic case, enjoy greater post-synaptic gain, and so have greater influence over the unfolding regime. High-weighted predictions, meanwhile, are less open to revision by sensory information, which would arrive (in the usual PP fashion) as prediction errors[11]. In my own treatments, precision-weighting is the secret of the putatively happy union between PP and work in EEE (embodied, extended, enactive) cognition. This is because varying precision estimations vary the balance of power between incoming signals and top-down predictions, and also create transient webs of effective connectivity. They thus deliver what is in effect a succession of special-purpose wiring diagrams for the brain, with radically different diagrams reflecting different tasks and different (bodily and environmental) contexts. Moreover, those special-purpose wiring diagrams are responsible for generating actions that (in context) exploit environmental opportunities and harvest new sensory information- a process that results in new swathes of predictions and precision-estimations. Brains like that are best seen as potent nodes in a rolling process of self- reconfiguration: one that weaves brain, body, and worldly opportunities into a succession of transient but highly efficient problem-solving wholes.

It is variable precision-weighting that thus allows us (Clark (2016a) (2017)) sometimes to look very much like well-tuned ecological pick-up devices, that simply latch onto whatever simple environmental cues will best guide ongoing action in the world, and at other times more like reflective engines, pondering our next move from the deepest reaches of our generative world-model. In the former case, a transient web of precision estimations ensures that task-relevant prediction error is rapidly and efficiently dealt with using sparse resources. In the latter case, prediction error penetrates deeper and deeper, requiring more

and more neural (and perhaps extra-neural) resources to damp it down to the tolerances of anticipated noise. I have described (Clark (2016a)) poster-child cases such as running to catch a fly ball in baseball in essentially the former terms, but any skilled thinker or performer is adept at spinning transient webs of connectivity that enable her to solve complex puzzles using minimal resources. As I often remark, the most under-appreciated role of a *rich* world-model is identifying situations in which a slimmer fragment is all that is needed to do the job.

Hohwy then raises an important puzzle. On the one hand, I'm committed to depicting much of human (and non-human) performance as rooted in these fast, efficient, strategies. This is one reason why I don't see PP as an insulating, intellectualist, internally-reconstructive approach. On the other hand, there is no doubt that the careful spinning of these webs of delicate transient, world-engaging connectivity is itself a fairly high-grade achievement. There is, as he puts it (this volume, ms p.4) "a potential tension here between both allowing and withholding a role for rich models".

The tension is especially marked for Hohwy, since he argues that we must repeatedly *infer* when we are in situations where low-cost solutions are viable, and that that requires (op cit p.5) "continuous modelling and tracking of *all* the relevant potential causal interactions across all contexts". Indeed, Hohwy goes further still adding that "predicting that there will not be any volatility-inducing causal interaction in a given context requires just as much rich modeling as predicting that there will be interactions" and that "without the rich model actually and continuously exerting its influence even in the conditions suitable for quick and dirty processing, there is no principled…setting of the gain on the prediction error".

But I wonder if this isn't a subtly mistaken picture of the development and application of expertise? It seems entirely possible that what training and immersion eventually deliver is the capacity to use simple environmental cues to activate (in the absence of any higher-level defeaters) a web of *default* precision assignments that install the transient organizational structure that best confronts that kind of puzzle in that kind of context. The concert pianist, sat at the piano, in the familiar concert hall, would thus fairly automatically recruit a web of precision estimations apt for the task. Were some novel circumstance or new instruction to intrude,that would generate prediction errors whose resolution recruit additional resources. But in the ordinary unfolding, I see no reason to think that the richness of the full world-model need actively be in

play. So I am unpersuaded that my appeals to variable precision-weighting merely re-invoke the full rich world-model 'one level up' as it were.

This is not to deny that there really is, in advanced minds, what Hohwy describes as "immense storage of causal knowledge". There surely is. But moment-by-moment, self-organizing around the computable quantity of prediction error, we manifest as a succession of relatively special-purpose brain-body-world devices, strung together by those shifting but self-organizing web of precision-weighting. In Humean spirit, no ponderous, all-knowing homunculus sits atop this web, carefully deciding moment-by-moment just what to do, and how to assign precision. Self-organizing around prediction error delivers precision variations, hence new effective connectivity patterns. These drive actions that weave the multi-level prediction machinery deeply into our worlds. We are perhaps misled because in the human case there sometimes emerge transient special-purpose devices for reflecting on our own nature, and on what kinds of future devices we wish to usher into being. One day, perhaps, we will see this as just 'more of the same', finally exorcizing the Cartesian demon from the tapestry of mind.

Orlandi and Lee set out to resist a certain gloss on PP. The gloss, in their words, depicts PP as "essentially a top-down, expectation-driven process, on which perception is aptly thought of as "controlled hallucination"" Orlandi and Lee (this volume) ms p.1. Pushing back against this kind of gloss, Orlandi and Lee write that:

> "in a novel environment, at least initially, the visual system's priors will be neutral between many possibilities, and the bottom-up signal will do most of the work. [So] there's nothing in the model ruling out current evidence often being much more informative than past evidence (….contrary to Clark's emphasis)." Orlandi and Lee ms p.4

There are at least two issues being raised here. One concerns the mechanics of dealing with novel environments. The other, more subtle but I think ultimately more interesting and important, is their subsequent claim that sometimes, even assuming the PP story, the bottom-up signal thus does 'most of the work'. This is true in one way but false in another, as we'll shortly see.

Let's start, though, by looking at the mechanics of perception in a novel environment. Imagine that I am taken from my bed one night, sedated, and then awake in a brand new, wholly unexpected, place. How do I recognize what kind of place I am in? Here is the PP story as told by Barrett and Bar (2009),

and rehearsed in Clark (2016a) p.42). First, very general, extremely rapidly processed (low spatial frequency) features of the sensory input enable an initial guess at the rough gist of the scene - is it a natural scene, a face, animals, an industrial landscape…? With apt gist-level prediction active, the full apparatus of top-down prediction gets a grip, as flurries of finer and finer-grained predictions concerning the details of the scene are generated and tested against the sensory evidence. The emergence of a rich and stable percept thus depends heavily, even in this unusually extreme case, upon the apt flow of top-down prediction, even though it is instigated using early rapid processing of low-level sensory cues.

But we should also ask how often we find ourselves in such truly unexpected environments anyway? In the ecologically normal run of things, I am not often kidnapped, nor is my brain suddenly 'turned on' in some fundamentally unpredicted situation. Instead, my life mostly consists in moving through a succession of quite substantially predictable environments, many of which I actively bring forth. For example, I find myself in the shopping mall because I set out to acquire a new set of headphones. As I move through this actively self-solicited space, multiple apt predictions are at all times in play, some of which meet resistance from the world, generating prediction errors that select, tune and nuance the flow of prediction[12].

But all this is really just a skirmish around the edges of a much deeper question, that concerns what it means to say that sometimes the bottom-up signal still 'does most of the work'. Heard one way, this is clearly and importantly true. Much of the power of the PP schema lies, as we have repeatedly seen, in the use of variable precision-weighting to modify the relative impacts of predictions and prediction errors according to self-estimated uncertainty. That immediately implies that under some circumstances perception can be very powerfully driven by the incoming sensory signal, while under others (such as viewing the hollow mask illusion) our predictions trump many aspects of the incoming signal. But there remains a sense in which, from a PP perspective[13], giving high-weighting to the sensory evidence is not quite the same as having the bottom-up flow do 'most of the work'. This is because, contrary to Orlandi and Lee, the top-down and bottom-up flows are not symmetric. Top-down predictions interact with each other in highly complex (non-linear) ways, while bottom-up prediction errors (PP suggests) do not.

To take a simple example (from Adams et al (2013) p.613) my predictions of the local sensory flux turn upon knowledge about objects, but any given object may be partially occluded by other objects in the scene. Predictions thus *weave*

together, reflecting all that we know about causes and complex inter-dependencies in the world. Prediction errors, meanwhile, are free to behave in much simpler ways, carrying their bespoke residual information until some complex weave of prediction quashes it – or occasionally surviving to drive plasticity and contribute to long-term learning. In this way descending predictions respect complex non-linearities within the generative model, while ascending errors do not.

Orlandi and Lee are thus wrong to conclude (ms page 7) that "we are… just combining two separate estimates of the stimulus [one top-down and one bottom-up] in a way that is essentially symmetrical between processing directions". For all the heavy (non-linear) lifting, PP asserts, is done by the predictions. This crucial functional asymmetry provides the deep reason why the PP framework, *despite* allowing for highly variable balances of power between sensory evidence and top-down prediction, is nonetheless properly described as one in which the top-down flow does special (and especially powerful) work. This also bears on their observation that "the right interpretation of the predictive coding model is that it does involve information about stimulus features being fed forward". In one sense this is uncontentious. Prediction errors are information (mathematically, they are the original information minus the prediction). So information about stimulus features are indeed being fed forward (though it is information relative to a prediction). But this, for the reasons just scouted, does not render the top-down and bottom-up flows symmetric.

Orlandi and Lee end by raising important questions concerning the relative merits and demerits of optimal feedback control (OFC) and active inference. Some of these turn on quite technical issues that are beyond the scope of this reply. But I would want to resist their suggestion that OFC is somehow *more explanatory* just because active inference replaces cost functions with predictions. The choice of cost functions (e.g. for a mobile robot) is itself an infamously black art. At worst, active inference replaces that black art with another – the installation of the priors that enable apt prediction. But in fact this reallocation (in which cost functions are treated as priors) has many useful consequences (rehearsed in Clark (2016a) chapter 4). For example, it is known that everything that can be specified by a cost function can be specified by some prior over trajectories, but not vice versa (see Friston (2011). Related concerns have led both dynamicists and roboticists to argue that explicit cost-function based solutions are inflexible, make unrealistic demands on online processing, and lack biologically plausible means of implementation (Thelen and Smith (1994), Feldman (2009), Mohan and Morasso (2011)).

Which brings me to remarkable and illuminating contribution by **Jesse Prinz**. Prinz asked, in effect, what (if anything) held my various views together? What agenda, proclivity, or persisting take on the mind provides the common ground between my work on connectionism, robotics, embodied and extended cognition, and predictive processing (PP)? My self-narrative, Prinz notes, has it that PP at last provides the single, overarching umbrella that neatly encompasses the best insights from them all. I do believe that to be true. But the question Prinz poses runs deeper. It runs deeper in roughly the way that insights delivered by various forms of psychoanalysis run deeper than the behavior patterns themselves, even if they fit together neatly enough. Indeed, by the end of Prinz' generous yet probative treatment, I felt very much as if I had been through a deep and highly personal process, fallen predictably in love with my analyst, and learnt something new and useful about my own goals and motivations. By way of reply, I'll simply re-trace his route, offering my own reactions (from the couch, as it were) along the way.

Prinz starts by noting some themes from the early work. *Microcognition* (back in 1989) was aiming to preserve many insights (especially those concerning structured information processing) from classical AI while at the same time defending a deeply connectionist picture of core biological processing. By 1997, with *Being There* and work on the extended mind, I was looking hard at how richly structured and (especially) richly self-structured environments enable us to press ever grander results from that associative, pattern-obsessed core. In the noughties, *Natural-Born Cyborgs* and *Supersizing the Mind* offered a whoselsale picture of human nature as one whose USP is the blurring of boundaries between mind, body, and world. All that was followed by *Surfing Uncertainty* with its picture of our biological brains as (embodied, situated) multi-level prediction engines. How do these pictures inter-relate?

One tempting option, Prinz suggests, is 'evolution'. Perhaps each picture slowly adds to and refines the one before it, while themes of dynamic, de-centralized, scaffolded adaptive response recur throughout. But there are apparent disconnects too. Am I meaning to stress external scaffolding structures or internal predictions? Is cognition a bag of tricks or a single unified process? Can predictive minds really extend? Many of the other pieces in this volume pick up on this kind of question, asking - for example - if I mean to depict us as rich modelers or fluent ecological couplers (both, of course!).

Prinz suggests a potential gulf between connectionism and the picture of the predictive brain. I'm not convinced by this. Historically, core aspects of PP

emerged straight from a connectionist lineage via the work of Geoffrey Hinton and others – see e.g. Hinton (1990) (2007). But more importantly, I don't think that PP involves a departure from a de-centralized dynamic vision, as Prinz seems to fear. On the contrary, PP strikes me as the best version yet of just such a vision. Prinz is concerned that PP may invoke 'over-arching monitors that set values for precision' (Section 3). That idea re-appears a bit later, when he wonders how the decision between rich model-based response and shallow more 'model-free' strategies is itself made, suggesting that "it is not clear how we make decisions about which strategy to deploy " then asking "Who is the "we" here? How do we manage the computational cost of such meta-cognitive oversight?". Just as Prinz notes, I am optimistic here. This is because precision estimations are both learnt and later recruited in exactly the same way as predictions, by self-organizing around prediction error signals. As a result, PP is highly compatible with the de-centralized dynamic vision from the early work.

Prinz is also doubtful that PP plays well with extended cognition, suggesting that if prediction is the mark of the mental, notebooks and their ilk look doomed to remain outside the cognitive apparatus. Here though, we need only note that those waves of precision-weighting weave inner predictions and world recruiting actions into a single problem-solving web, while also providing what was long missing – a principled account of how the inner and outer resources get together in just the right ways at just the right times. Perhaps prediction is indeed the mark of the *terrestrial bio-mental* – but nothing in the arguments for the extended mind required that there be no common core to the bio-mental contributions.

Still, Prinz is right: none of these frameworks are quite the same and PP delivers, as he so delicately puts it, some quite serious 'perturbations'. Unlike connectionism, PP systems depict a deep functional asymmetry between downward-flowing predictions and upward-flowing errors. And unlike some work on embodied cognition, PP relies heavily on structured inner knowledge. Differences could be multiplied.

It is at about this point that Prinz dons a more 'philosophy of science' hat, suggesting that although my later positions cover many of the same cases, and may even use many of the same terms as before, the *meanings* of key terms may now shift and alter, in ways that subtly transform their empirical content and may even impact their metaphysical implications. Representations, for example, may be vectors, attractors, filters, or forward models. So even what looks like a seamless merging of older and newer ideas and perspectives may well involve revisions that strike deep into the heart of the earlier pictures.

What came next was, for me, a striking (almost psycho-analytic) moment of intellectual self-discovery. For each new framework, Prinz then argued, is really offering a new way of *seeing ourselves*, making everything spin, morph, and alter, around some new central construct, be it distributed representation, coupling, de-centralized control, offloading, or predictions. What I am selling, then, is really a succession of ways of understanding the mind – distinct perspectives each of which puts something new and potentially transformative at the center of the cognitive universe.

That could be worrying. How could it be good to be so fickle? It is then that Prinz serves up a lovely and comforting thought. Could it be that each perspective is simply a kind of invitation to try out a new way of seeing ourselves, and the place of mind in the universe? Perhaps no one such picture can really hope to deliver the whole truth. But by fully inhabiting each picture, we learn something about who and what we are, and what we can hope to be. The exercise, Prinz suggests, is as much one of art as science. It is about bringing some stuff to the fore, and asking the reader to look at the world (even if only temporarily) through that lens.

That rang true. It gave me a new way to look at my own work: a new way to think about both the changing landscape and my abiding sense of a continuous, progressive story. Prinz's contribution thus performs exactly the function he so generously offers for mine – it changed my manifest reality, giving me new ways to see and inhabit my world[14].

Anil Seth's compelling contribution directly confronts core aspects of manifest human reality, using the rich resources of interoceptively inflected PP. Interoceptively inflected PP, Seth argues, moves us away from the picture of minds as mirrors (faithful, accurate inner models) of nature. Instead, we are invited to embrace our nature as 'beast machines' dedicated most fundamentally not to mirroring but to *control* – both the control of action and of our own inner states, so as to keep our systemic organizations within viable bounds. And therein, Seth argues, may lie the origins of feelings of embodiment, selfhood and subjectivity, and the key to understanding the emergence of consciousness in the material realm.

The link between control and prediction is pretty direct, since to keep complex systems within bounds often requires anticipating breaching those bounds and taking pre-emptive steps to remedy the slide before it becomes too late. Indeed, Clark and Grush (1999) argued, pre-PP, that this kind of pre-emptive predictive

control marks the very spot at which genuine representing first appears in the natural order, as systems develop *forward models* able to support fluent action despite substantial delays in corrective feedback from the peripheral motor plant. What thus holds true for gross motor action is doubly true, as Seth powerfully notes, for internal 'actions' that support homeostatic regulation. Creatures must pre-emptively correct before blood sugar levels become too low, or before body temperatures reach below or above certain bounds – a process known as allostasis rather than simple homeostasis. Seth (and see also Friston (this volume) suggests that our neurobiological nature as prediction machines is rooted in this need for allostasis.

Interoceptive predictions, Seth and others have argued[15], help construct feelings and emotions by entering into Bayesian inferences that integrate information about context, action, and our own bodily state. These inferences combine exteroceptive, proprioceptive, and interoceptive information, issuing in predictions that engage the world and entrain the body. But this leaves a puzzle: How, Seth asks "can PP account for the qualitative differences between perceptual experiences of the external world, and self-related experiences such as emotion and experiences of having, and being, a body?". These different experiences will each involve bringing together a host of information sources combined within a predictive matrix that self-organizes around prediction error. But they appear very different when viewed 'from within'. Emotions, unlike tomatoes, do not appear to be located in space, and feelings of body-ownership seem even less 'object-like' than those of emotion, going way beyond just knowing how and where conjoined body-parts body are located in space. Seth argues that that interoceptive signals play an especially large role such cases, and that "non-object-like phenomenology is linked to the control-oriented nature of interoceptive predictions."

In line with 'sensorimotor contingency theory' (O'Regan and Noë (2001)), Seth (2014)) suggests that our sense that the tomato is a solid object located in space may be linked to webs of sensorimotor know-how that involve predictions of how it would look and behave were we to act upon it in various ways, or to move around it. It is these predictions that constitute the experienced content of the tomato (unlike the emotion, say) being a solid 3D object in the world. The content of the percept is thus strongly influenced by the web of exteroceptive and proprioceptive (action-guiding) predictions. This, in turn, leads Seth to argue that these perceptual contents depend more on predictions than prediction errors since these key counterfactual predictions cannot (while remaining counterfactual) generate any prediction errors at all. But

interoceptive predictions behave differently, or so Seth argues. Actions rooted in our command of sensorimotor contingencies reveal counterfactually predicted facets of the external world. But inner allostatic actions change the inner world to fit predictions about its proper shape and bounds. They are control-oriented rather than discovery-oriented. Perhaps this difference explains the non-object-like phenomenologies of mood, emotion, and the physical embodiment?

This is a neat idea, and one that I think has much to recommend it. It seems unlikely, though, that we here confront a very firm divide so much as a graded and changeable balance, in which nearly every conscious experience is interoceptively, exteroceptively, and proprioceptively informed in ways that vary with task and with inner and outer context. There may be puzzle cases too, such as the experience of pain. Seth (2013) mentions pain, but it's not clear to me how well it fits the model proposed. Some pains (sports injuries especially) respond quite well to self-directed actions, such as rubbing or ice-ing, while others, such as organ pain, seem locked into place in ways that no self-directed action systematically impacts. In the former case (only) we soon learn a swathe of pain-related sensorimotor contingencies. But it is not clear which of these scenarios (if either) corresponds to experiencing the pain as more or less object-like. My own intuition is that it is the pain that is *not* subject to counterfactual motor-based manipulations that seems most object-like, potentially providing at least one case where the phenomenology does not follow the pattern Seth suggests.

Seth ends with an eloquent description of how our nature as organic, self-preserving 'beast machines' underpins our sense of embodied being in the world by placing interoceptively-informed predictive control at the core. Does this kind of picture makes real progress with the mystery of consciousness itself? I think it does. One thought emerging from our current ERC-supported project ('Expecting Ourselves'[16]) is that PP invites us to theorize conscious experience within a kind of 3-space whose axes are (i) the depth of the generative model (ii) the context-varying inflection of that model by interoceptive information, and (iii) the degree of self-modeling present (where that speaks to the sense of future-oriented unfolding and personal narrative that emerges when creatures that already score on the other two axes add themselves as new nodes or latent variables within their own generative model). It is that last dimension that allows some (but only some) sentient creatures to explore increasingly complex counterfactual scenarios contingent upon their own actions and choices. We suspect that it is also that last dimension that leads such creatures to find qualia and conscious experience especially puzzling.

For they model not just the external world, but *themselves*. Such beings belong to an intriguing group of systems that 'expect themselves', inferring that they are persisting agents that experience pains, see red, react to dangers, and encounter tempting tasty treats when opening fridge doors.

Much of our scientific and philosophical puzzlement about 'qualia' may thus turn out to be misplaced, reflecting not some strange ontological status for qualia but our failure to appreciate that we model ourselves and our own reactive dispositions in much the same (simplified, compressed, pragmatic) way as we model the external world. We are beings that model ourselves as having qualitative experiences. Understanding the self-modeling origins of our own puzzlement concerning subjectivity and experience, while recognizing the many ways interoceptive and exteroceptive information combine in the service of adaptive success, may be the best recipe for solving the so-called 'hard problem' of consciousness itself (for forays into this territory, see Dennett (2013) (2015), Clark (2016b)).

We humans score highly on all three dimensions of this 'consciousness matrix'. But it is easy to imagine other creatures and systems that score low on all three, or that score highly but only on one or two of them. A game-playing agent of the kind described by Mnih et al (2015) may have a deep (though very special-purpose) generative model, but one that is not inflected by interoceptive information, and one that moreover operates without need for rich self-modeling. Some animals, by contrast, may have deep and multi-purpose generative models, richly inflected by interoceptive information, but without needing rich models of *themselves* as individuals with distinct purposes and goals. And a simple robotic agent might already manage a low score on all three axes. In all these cases, rather than ask 'so is that a conscious being or not' perhaps we ought (as suggested by Sloman (2010)) merely to note the pattern of similarities and differences from our own case?

Looking hard at living organizations quite different from our own is a powerful strategy, masterfully deployed by **Barabara Webb** in the closing essay. Webb specializes in understanding and modeling (often using robotic platforms) the adaptive strategies of insects - life-forms far removed from philosophy's mammalian stomping grounds. Such life-forms, Webb rightly suggests, provide powerful test-cases for theories such as PP, insofar as they may aim at truly unifying pictures of perception, cognition, and action.

Webb's earlier work figured prominently in the push towards treating mind as a richly embodied phenomenon, revealing (for example) the many ways in which

insect performance emerged from delicate and often unexpected combinations of gross morphology, action, and neural circuitry. That work, as she notes, put some pressure on unreflective appeals to rich inner models and encodings, since insect success was often rooted in specialized, embodied strategies that sidestepped the need for accurate general-purpose 'reconstructive' representations of distal causes. My hope, as Webb notes, is that PP now provides a framework deeply amenable to that same push, since it treats perception, cognition, and action as locked in a circular dance, self-organized around predictions, precisions (reflecting self-estimated sensory uncertainty), and prediction errors.

But does PP actually get a useful grip on the insect case? The evidence is interestingly mixed. On the plus side, Webb points to exciting recent work that suggests that crickets, fruitflies, and dragonflies all use forward models to issue delicate and accurate predictions of their own upcoming behaviors, enabling them to factor out changes caused by their own actions from other sources of sensory perturbation (Webb (2004), Kim et al (2015), Mischiati et al (2014)). There is also evidence of prediction-error driven learning, and a plausible implementation story for insect forward models using the mushroom body neuropil, which boasts a layered, orderly, feedback rich architecture. Webb also notes that cockroaches, crickets, and fruitflies also seem to navigate using 'place memory' encodings akin to those supported by the hippocampus in rats and humans. In the light of this, it is interesting to add that recent work suggests that hippocampal place cells "do not encode place per se but rather a predictive representation of future states given the current state." (Stachenfeld et al (2017)).

Forward models and prediction-error driven learning are thus plausibly in play, and active in multiple living organizations, including the insects. When it comes to the use of probabilistic generative models that deliver predictions in ways nuanced by self-estimated sensory uncertainty, however, things look more complicated. Webb's group showed (Wystrach et al. (2015)) that certain ants seemed to perform optimal cue integration, combining information in ways that reflect ideal Bayesian estimations of how to weight different cues at different times. Are the ants full, if tiny, PP agents, actively estimating their own context-varying uncertainty, and using that estimate optimally to orchestrate the combination of different sources of information? Perhaps. But another possibility, suggested by further experimental manipulations, is that the ants moving away from the nest use a simple proxy (home vector length) that stands in for self-estimated sensory uncertainty. This easy-to-compute quantity

acts as a special-purpose trick that mimics – for the kind of cue integration the ants actually need – the calculation of their own path-integration uncertainty.

This leads Webb to raise an important question: "If the ant's ability to do optimal cue integration has been hard coded by evolution, does it still count as evidence for the PP view?" Staunch proponents of the full 'free energy minimizing' story may say that it does, while noting that in this simple case evolution has settled for a hard-coded and very special-purpose approximation. My view, however, would be that the ants, if they are indeed using only this special-purpose proxy for self-estimated sensory uncertainty, are not fully-fledged PP agents. Instead, they are the kinds of interesting mixed case one might expect to find in a wide range of simpler beings. They are equipped with core elements (forward models and error-based updating) of the predictive apparatus, but are not yet able to leverage their own acquired knowledge in an open-ended variety of different contexts using the kinds of flexible precision-weighting apparatus found in higher animals (and implemented using a wide variety of resources including dopaminergic modulation and phase-locked neuronal oscillations).

Does that put pressure on the claim that PP is a truly unifying story? I don't think so. For as we saw earlier PP, unlike the full free energy story, does not claim to apply to every form of living being. Instead, PP describes a specific, powerful, and amazingly flexible mechanism that represents one solution to the more general problem of temporarily resisting (or rather, appearing temporarily to resist!) the second law of thermodynamics. Considered as a process theory of that type, it is unsurprising if it turns out that the evolutionary path to the full apparatus includes many stepping-stones, fragments, and outright alternatives. This is consistent with the idea that core principles underlying life and mind involve the use of hard-coded or more flexible means of predicting the sensory flow, and (hence) that the free energy minimizing story helps us appreciate the continuity between basic life and advanced minds. If nothing else, the PP story thus invites us to ask new questions of nature, and may suggest new ways of gradually accumulating the resources that deliver general-purpose problem solving and rich conscious experience of a structured world.

What about the promised union between embodied cognition and PP? Is that threatened by the discovery, in ants (and perhaps sometimes in humans too) of special-purpose tricks that sidestep the need to estimate sensory uncertainty? This would be the case only, it seems to me, if it tuned out that humans *always* rely on special-purpose tricks rather than *also* benefitting from a general-purpose trick (variable precision-weighting) that enables us to make flexible

use of the vast body of acquired knowledge realized in our probabilistic generative model. Our remarkable capacity to at least appear to be general-purpose problem-solvers may be rooted, it seems to me, in that key multi-purpose trick: a trick that then gets amplified first by spoken language (which arguably provides an additional means of manipulating our own precision-weighting apparatus – see Lupyan and Clark (2015)) and then again by the strategic use of social and material culture (see Roepstorff (2013), Roepstorff et al (2010)).

Finally, notice that general-purpose and special-purpose tricks and ploys may then nest and iterate in complex and extremely powerful ways. With the general –purpose precision-weighting trick in place, new special-purpose tricks may be learnt and deployed wherever task and context allow. That kind of nesting could extend all the way to the learning and use of cheap proxies for our own self-estimated uncertainty! We might, that is to say, learn to use precision-weighting variations to install transient patterns of influence that allow some simple quantity to stand in for more complex (though always merely approximate) online calculations of uncertainty whenever task and context permit. This would be just one more case in which the presence of rich inner resources is what enables an efficient embodied strategy to emerge and carry the load. Variable precision-weighting, channeled and directed by language, culture, and material scaffolding, may be what enables our own stunning combination of deep flexibility and efficient embodied problem-solving - the signature of mind incarnate.

Acknowledgments

This Reply was written thanks to support from ERC Advanced Grant XSPECT - DLV-692739.

References

Adams et al (2013 - Predictions not commands: active inference in the motor system *Brain Struct Funct* (2013) 218:611–643

Barrett, L. F., & Bar, M. (2009). See it with feeling: Affective predictions in the human brain. *Royal Society Phil Trans* B, 364, 1325-1334

Barrett, L. F., and Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, 16, 419–429.

Bechtel, W. (1994) Natural deduction in connectionist systems. *Synthese*, 101, 433-463.

Bechtel, W. (1996) What knowledge must be in the head in order to acquire language. In B. Velichkovsky and D. M. Rumbaugh (Eds.), *Communicating meaning: the evolution and development of language*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bruineberg, J., & Rietveld, E. (2014) Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in human neuroscience*, 8 , 599.

Calvo, P., & Friston, K. (2017). Predicting green: really radical (plant) predictive processing. *Journal of The Royal Society Interface*, *14*(131)

Chalmers, D (2005) The Matrix as Metaphysics. In C. Grau (ed) *Philosophers Explore The Matrix* (Oxford University Press, NY)

Clark, A (2005) "The Twisted Matrix: Dream, Simulation or Hybrid?" in C. Grau (ed) *Philosophers Explore The Matrix* (Oxford University Press, NY)

Clark, A (2008a) *Supersizing the Mind: Action, Embodiment, and Cognitive Extension* (Oxford University Press, NY)

Clark, A. (2008b). Pressing the flesh: A tension in the study of the embodied, embedded mind? *Philosophy and Phenomenological Research*, 76 (1), 37–59.

Clark, A. 2009a. Letter to the editor. London Review of Books 31:6 (26 March 2009).

Clark, A. 2009b. Spreading the Joy? Why the Machinery of Consciousness is (Probably) Still in the Head. *Mind* 118:963-93.

Clark, A (2013) Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science *Behavioral and Brain Sciences* 36: 3: p. 181-204

Clark, A (2015) What 'extended me' knows. Synthese, 192(11), 3757–3775.

Clark, A (2016a) *Surfing Uncertainty: Prediction, Action, and the Embodied Mind* (Oxford University Press, NY)

Clark, A (2016b) Strange Inversions: Prediction and the Explanation of Conscious Experience *Engaging Daniel Dennett*. Huebner, B. (ed.). Oxford University Press

Clark, A (2017) Busting Out: Predictive Brains, Embodied Minds, and the Puzzle of the Evidentiary Veil. *Noûs*, 51: 727–753.

Clark, A., and Grush, R. (1999). Towards a cognitive robotics. *Adaptive Behavior*,

7(1), 5–16.

Colombetti, G. & Roberts, T. (2015). Extending the extended mind: The case for extended affectivity. *Philosophical Studies*, 172 (5), 1243-1263.

Davidson, Donald (1987)). "Knowing One's Own Mind" *Proceedings and Addresses of the American Philosophical Association*, 60 (1987), 441-58.'

Dennett, D. C. (1978). Where Am I? In *Brainstorms: Philosophical Essays on Mind and Psychology* (pp. 217–231).

Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA, MIT Press.

Dennett, D. (1998). *Brainchildren: Essays on Designing Minds*. Cambridge, MA, MIT Press.

Dennett, D (2013) Expecting ourselves to expect: The Bayesian brain as a projector *Behavioral and Brain Sciences* 36: 3: 209-210

Dennett, D (2015) 'Why and How Does Consciousness Seem the Way it Seems?', in T. Metzinger and J.M. Windt (eds). *Open MIND:* 10(T), Frankfurt am Main, MIND Group, 2015, DOI: 10.15502/9783958570245.

Farkas, K. 2012. Two Versions of the Extended Mind Thesis. *Philosophia* 40:435-447.

Feldman, A.G. (2009). New insights into action-perception coupling. *Experimental Brain Research*, 194(1), 39-58.

Fodor, J (1986). Why Paramecia don't have Mental Representations. *Midwest Studies in Philosophy,* 10, 3-23.

Friston, K (2011) What is optimal about motor control? *Neuron* 72: 488-498

Graves, A., Wayne, G., Reynolds, M., …. and D. Hassabis. (2016) Hybrid computing using a neural network with dynamic external memory. *Nature*, 538: 471–476, 201

Heyes, C (2018) *Cognitive Gadgets: The Cultural Evolution of Thinking* (Harvard University Press)

Hinton, G. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence,* 46, 47–75.

Hinton, G. E. (2007a). Learning multiple layers of representation. *Trends in Cognitive Sciences*, 11, 428–434.

Kim, A.J., Fitzgerald, J.K. & Maimon, G., 2015. Cellular evidence for efference copy in Drosophila visuomotor processing. *Nature Neuroscience*, 18(9), pp.1247–1255.

Lupyan, G and Clark, A (2015) Words and the World: Predictive Coding and the Language-perception-cognition Interface. *Current Directions in Psychological Science* . 24, 4, p. 279-284

McDowell, J (1994) *Mind and World* (Cambridge, MA: Harvard University Press).

Miller, M., & Clark, A. (2017). Happily entangled: prediction, emotion, and the embodied mind. *Synthese*, 1-17

Mischiati, M. et al., 2014. Internal models direct dragonfly interception steering. *Nature*, 517(7534), pp.333–338.

Mnih, V., Kavukcuoglu, K., Silver, D., … Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*, 529.

Mohan V and Morasso P (2011) Passive motion paradigm: an alternative to optimal control. *Frontiers in Neurorobotics*. 5:4

Newell, A. & Simon, H. A. (1972) *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall

O'Regan, J. K. and Noë, A. (2001) A sensorimotor approach to vision and visual consciousness. *Behavioral and Brain Sciences* 24/5: 883-975.

Roepstorff, A (2013). Interactively human: Sharing time, constructing materiality. *Behavioral and Brain Sciences*, 36: 224-225

Roepstorff, Niewöhner, and Beck (2010) Enculturating brains through patterned practices *Neural Networks* 23: 1051-1059

Searle, J (1980) "Minds, Brains and Programs" *Behavioral and Brain Sciences* 1 : 417-424.

Searle, J (1992) *The Rediscovery of the Mind* (Cambridge, Ma.: MIT Press).

Seth, A K (2013) Interoceptive Inference, Emotion, and the Embodied Self. *Trends in Cognitive Sciences* 17, no. 11: 565–73.

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: explaining the puzzle of perceptual presence and its absence in synaesthesia. *Cogn. Neurosci.* 5:2: 97–118.

Seth, A.K, Suzuki, K., and Critchley, HD (2011) An Interoceptive Predictive Coding Model of Conscious Presence. *Frontiers in Psychology* 2:395.

Sloman, A (2010) An Alternative To Working On Machine Consciousness. *International Journal of Machine Consciousness* Vol. 2, No. 1: 1-18

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nat Neurosci*, *advance online*

Thelen, E and Smith, L (1994) *A Dynamic Systems Approach To The Development Of Cognition And Action* (MIT Press, Camb. MA)

Trewavas, A. (2003). Aspects of plant intelligence. *Annals of botany*, *92*(1), 1-20.

Turner, Scott J (2000) *The Extended Organism: The Physiology of Animal-Built Structures.* (Harvard University Press).

Webb, B. (2004). Neural mechanisms for prediction: Do insects have forward models? *Trends in Neurosciences* 27, 278–282.

Wystrach, A., Mangan, M. & Webb, B., 2015. Optimal cue integration in ants. *Proceedings of the Royal Society B: Biological Sciences*, 282(1816), p.1484.

[1] From the list of aphorisms published at:
http://www.cybsoc.org/ross.htm

[2] Chalmers notes that I make just that move in Clark (2009a). But as he later notes, my strategy there is really to use this as a softener for the main claim, which is that external operations can be woven in via perception and action and yet constitute an extended cognitive circuit.

[3] The case is thus similar to that of the long-term couple discussed in a footnote to the original paper.

[4] These devices are also known as Neural Turing Machines – see Hassabis (2017)

[5] This picture was first mooted in a never-published paper co-authored with Jesse Prinz back in the early 90's. That paper was entitled "The Absent Mind" and the core claim was that 'mind' and 'cognition' were confusing and unstable terms that should play no role in a mature cognitive science.

[6] The quote is from R. Martin's translation of his *De Anima* (*Liber de anima seu sextus de naturalibus*) vol 7 .

[7] Anderson and Chemero go on to discuss these issues in the context of some long-running debates concerning the implications of 'Markov blanket' organizations. I agree with most of their comments there, so won't pursue that angle here (see Clark (2017)).

[8] In the traditional Scottish ballad "The Bonnie Banks o' Loch Lomond.

[9] Thanks to Matteo Colombo for suggesting this way of looking at the difference.

[10] This paper – perhaps more than any other in the volume – raises many more questions than I can attempt to answer here. So I have chosen to focus on just one core and contested issue, the question of whether we can fruitfully combine the PP picture with more 'ecological' ones that stress frugal, efficient, non-reconstructive, modes of contact with the world.

[11] From a purely Bayesian perspective there is no difference between increasing the weight on prediction or decreasing the weight on (relevant) prediction error. But in the nervous system, these may well correspond to different manipulations, and the differences may matter, for the differentiation and treatment of psychopathologies.

[12] To be mostly moving through highly predictable spaces is also plausibly seen, as Friston (this volume) argues, as a necessary condition for our own existence. Staying within such bounds is the signature achievement of organizational forms that appear to mount staunch if temporary resistance to the second law of thermodynamics. What is here at issue, though, is something subtly different – whether remaining within those bounds involves, for creatures like us, the constant down-flow of predictions of their own sensory flux.

[13] Orlandi and Lee rightly distinguish 'merely Bayesian' stories from the PP process model. Still, it is clear from the outset that it is the gloss on PP as involving a characteristically "top-down, expectation-driven" process that they are seeking to call into question. They do this both by emphasizing the potential for strongly bottom-up, relatively stimulus- driven processing in hierarchical Bayesian settings, and by suggesting (in their section 2 discussion of PP proper) that we are just combining estimates in a way that privileges neither processing direction. The flaw in this argument turns (see text) upon a functional asymmetry at the core of PP– an asymmetry between the complex non-linear construction of downwards-flowing predictions and the much simpler forward-flowing accumulation of prediction errors calculated against that flow.

[14] Prinz himself does not work like this. He pursues topic after topic (emotions, concepts, consciousness) canvassing everything science, and sometimes art, has to offer on each individual front. Whose way is best? I don't think we need to choose. My old friend and mentor the cybernetician Donald T. Campbell insisted that the collective scientific endeavor is best served by pursuing a variety of different but overlapping and inter-communicating ways of exploring mind and its place in nature. He called this the "overlapping fish-scales" model of science and philosophy.

[15] E.g. Seth (2013), Seth et al (2011), Barrett and Simmons (2015), Miller and Clark (2017).

[16] See http://www.x-spect.org/